2012

Improving Biosurvillance: Optimizing Biosurveillance Systems

Fricker, Ronald D., Jr.

Invited speaker: "Optimizing Biosurveillance Systems," Quality and Productivity Research Conference, Long Beach, CA, June 2012.

NAVAL
POSTGRADUATE
SCHOOL

Improving Biosurveillance:
Optimizing Detection Thresholds

Ronald D. Fricker, Jr.

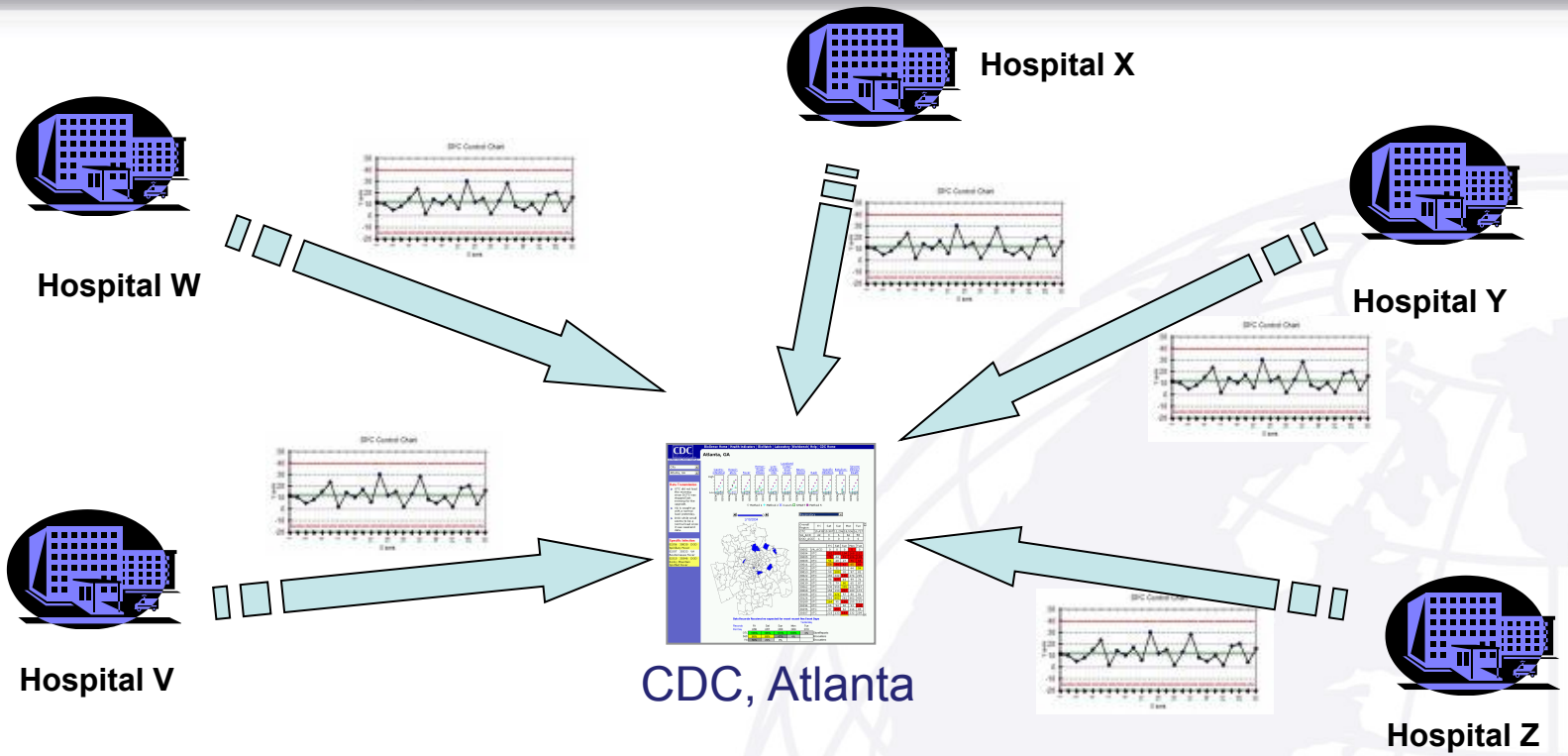
June 7, 2012

What is Biosurveillance?

- Homeland Security Presidential Directive HSPD-21 (October 18, 2007):
  - "The term 'biosurveillance' means the process of active data-gathering … of biosphere data … in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity." [1]
  - "The Secretary of Health and Human Services shall establish an operational national epidemiologic surveillance system for human health..." [1]
- Epidemiologic surveillance:
  - "…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." [2]

[1] www.whitehouse.gov/news/releases/2007/10/20071018-10.html
[2] CDC (www.cdc.gov/epo/dphsi/syndromic.htm, accessed 5/29/07)

Think of Biosurviellance Like a Large System of Shewhart Control Charts



- Issue: False alarms a serious problem
  - "*…most health monitors… learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!*" [1]

[1] https://wiki.cirg.washington.edu/pub/bin/view/Isds/SurveillanceSystemsInPractice

Formal Description of the System

- Each location sends data to system daily
  - Let X_{it} denote residual from model predicting counts from location i on day t
  - If no attack anywhere $X_{it} \sim F_0$ for all i and t
  - If attack occurs on day t at location i then $X_{it} \sim F_1$, $t = t, t+1, \ldots$
- Denote probability of attack at location i as p_i, where $\sum_i p_i = 1$
- Threshold detection: Signal on day t^* if

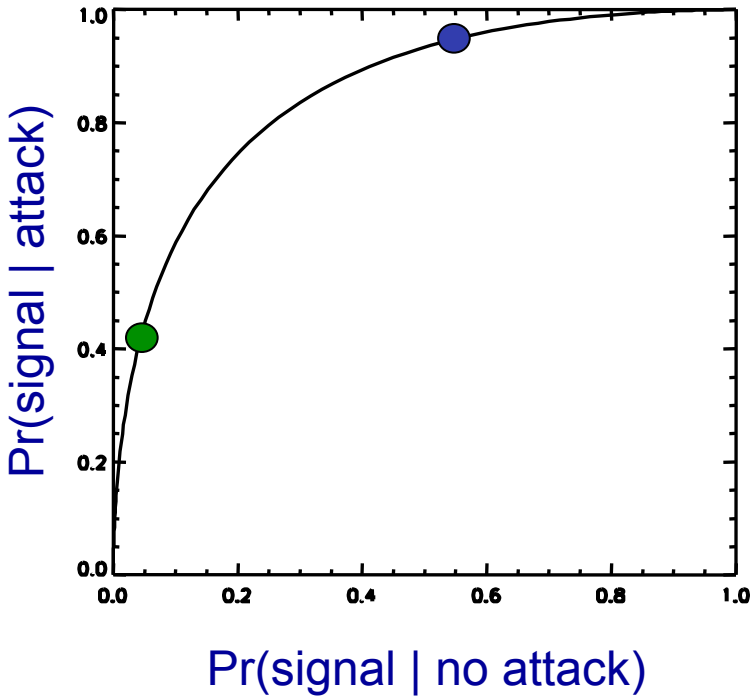
$$X_{it^*} \geq h_i$$

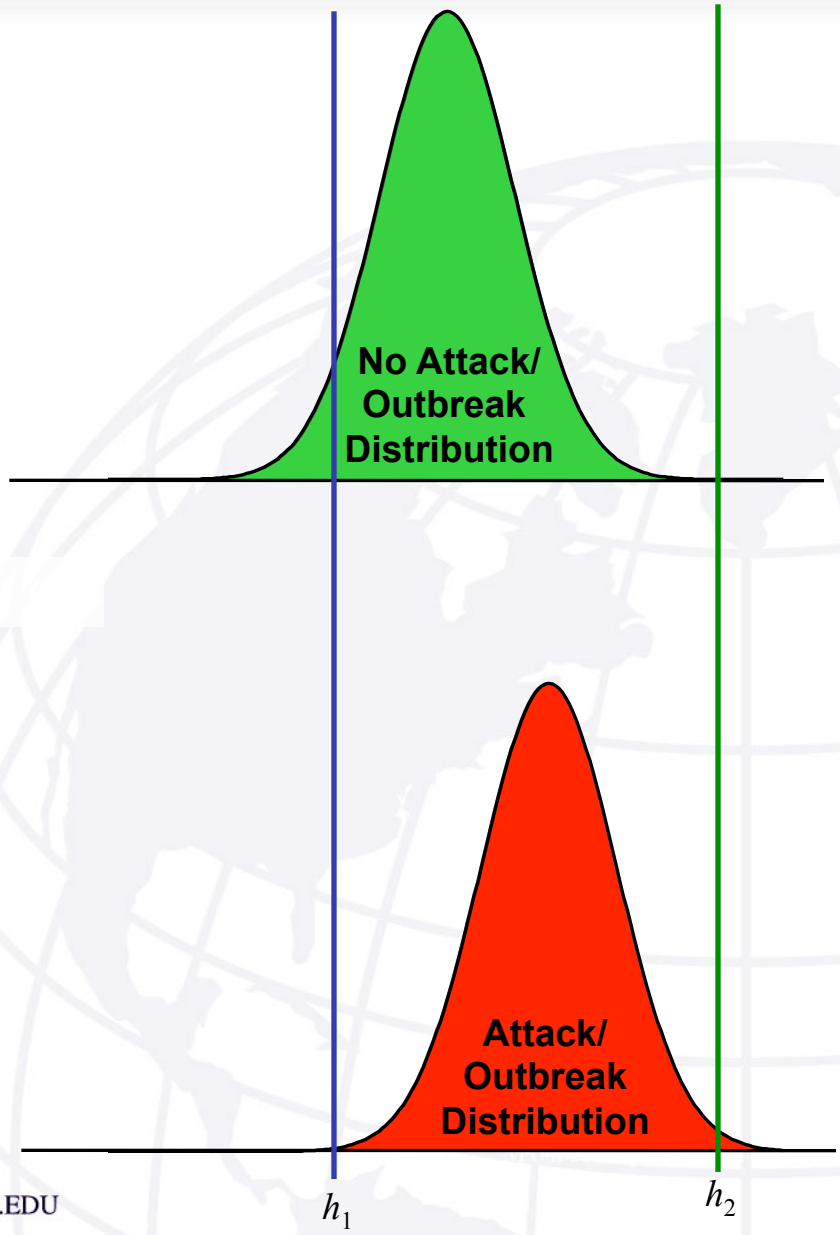
for one or more locations

It's All About Choosing Thresholds

- For each hospital, choice of h is compromise between probability of true and false signals

ROC Curve





Some Starting Assumptions

- Absent anomalies, the X_{it} and independent and identically distributed (iid) according to f_0
- If anomaly occurs, X_{it} iid according to f_1 for the affected data stream(s)

- That is, to start, we're assuming the observations are independent over time and between data streams
- To achieve temporal independence, may be monitoring residuals from model that accounts for systematic effects in the data

Mathematical Problem Formulation

- It's simple to write out:

$$\Pr(\text{detection}) = \sum_i \Pr(\text{signal}|\text{attack}) \Pr(\text{attack})$$

$$\text{E}(\#\text{ false signals}) = \sum_i \Pr(\text{signal}|\text{no attack})$$

- Express it as an optimization problem:

$$\max_{\mathbf{h}} \sum_i \left[1 - F_1(h_i)\right] p_i$$

$$\text{s.t.} \sum_i \left[1 - F_0(h_i)\right] \leq \kappa$$

An Illustrative Example

- Absent anomalies, (standardized) data distributed according to standard normal:

$$F_0 = \mathrm{N}(0,1)$$

- Anomaly manifests as a 2σ increase in mean:

$$F_1 = \mathrm{N}(2,1)$$

- Then, problem is:

$$\min_{\mathbf{h}} \sum_i \Phi(h_i - 2) p_i$$

$$\text{s.t.} \sum_i \Phi(h_i) > n - \kappa$$

- Let $n = 10$ with the following p_is:

Ten Hospital Illustration

Hospital i	p_i	Common Threshold #1	Optimal Threshold (h_i)	Common Threshold #2
1	0.797	2.189	1.068	1.310
2	0.064	2.189	3.602	1.310
3	0.056	2.189	3.732	1.310
4	0.048	2.189	3.915	1.310
5	0.013	2.189	4.656	1.310
6	0.006	2.189	4.736	1.310
7	0.006	2.189	4.736	1.310
8	0.005	2.189	4.755	1.310
9	0.003	2.189	4.773	1.310
10	0.002	2.189	4.791	1.310
	P_d	0.117	0.378	0.378
	$\sum \alpha_i$	0.143	0.143	0.951

Simplifying the Optimization

- Monitoring n data streams means optimization has n free parameters (thresholds)
  - Hard for to solve for large systems
- Constraint can be expressed as an equality
  - See Fricker & Banschbach (2012) for proof: http://faculty.nps.edu/rdfricke/frickerpa.htm
- Then can wrap the constraint into the objective function
  - Turns it into an unconstrained maximization problem
  - Unconstrained problem likely easier to solve

Specific Result Assuming Normality

- Assuming normality (and equal variances), can simplify to one-parameter problem:
  - *Lemma*: For F_0=N(0,1) and F_1=N(γ,1), the optimization simplifies to finding α that satisfies

$$\sum_{i=1}^{n} \Phi\left(\alpha - \frac{1}{\gamma} \ln(p_i) \right) = n - \kappa,$$

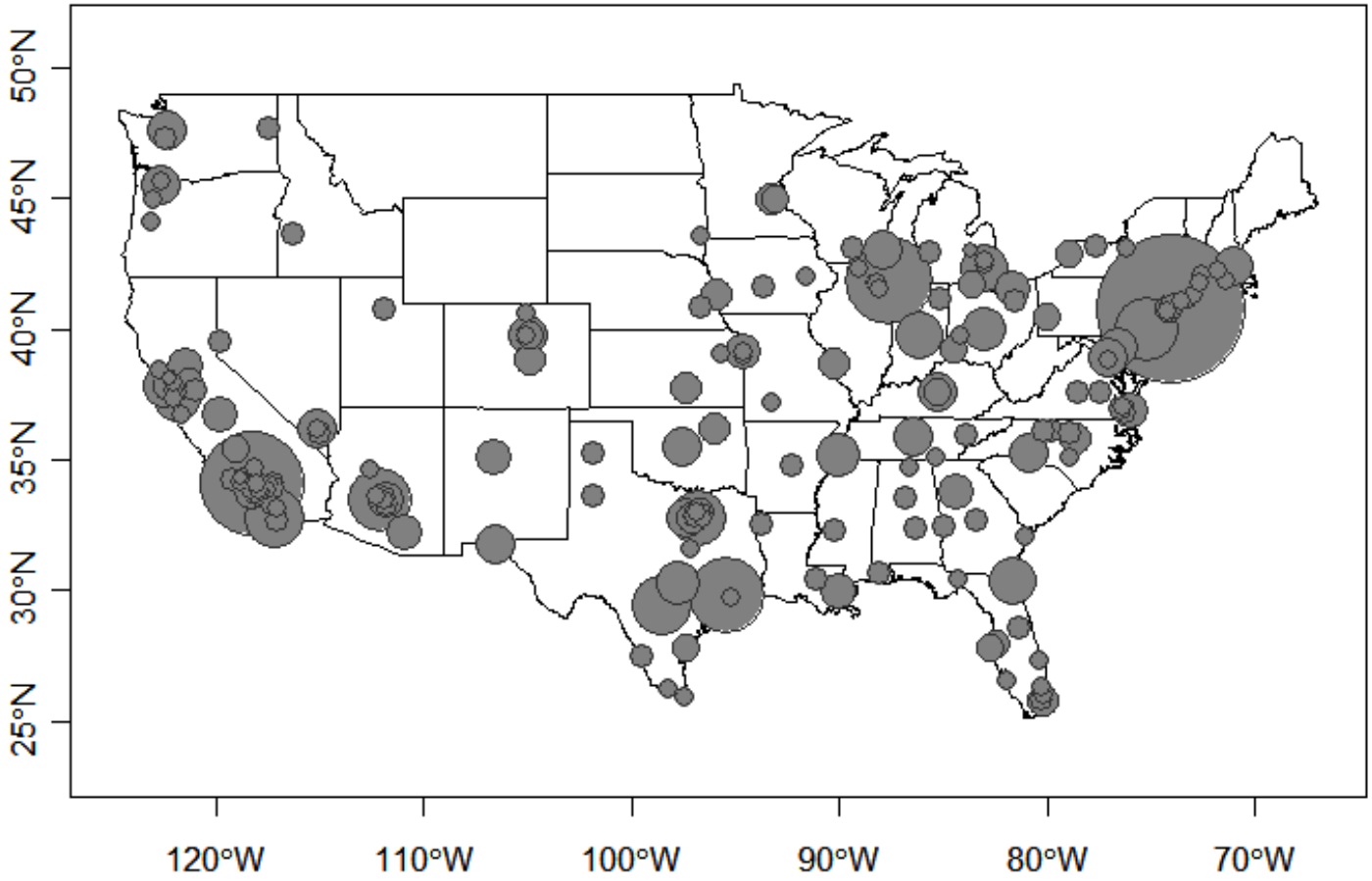
and the optimal thresholds are then

$$h_i = \alpha - \frac{1}{\gamma} \ln(p_i).$$

- See Fricker & Banschbach (2012) for derivation

Consider (Hypothetical) System to Monitor 200 Largest Cities in US

- Assume probability of attack is proportional to the population in a city



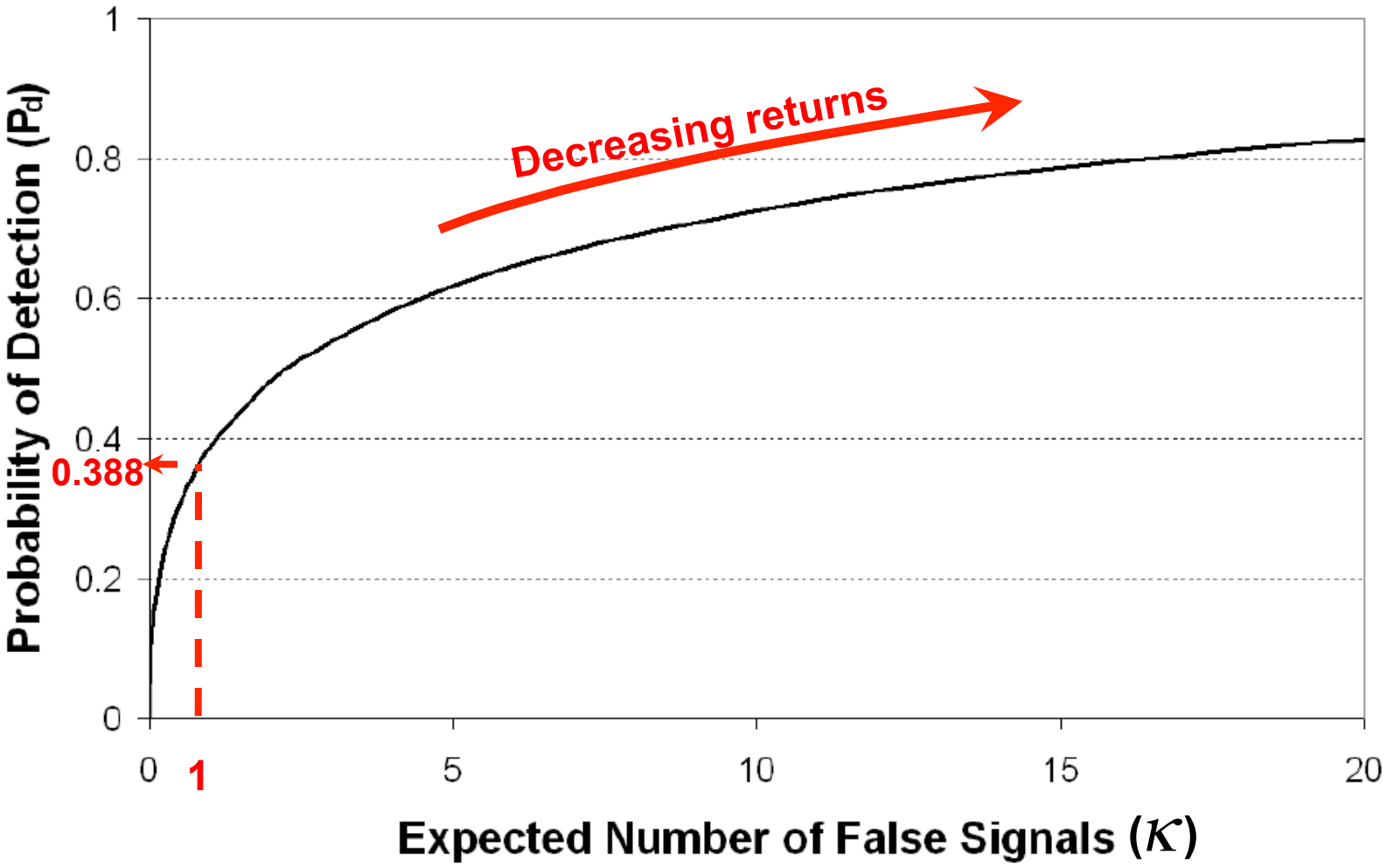
Optimal Solution for 200 Cities

- Assume
  - 2σ magnitude event
  - Constraint of 1 false signal system-wide / day

i	City	State	Population	Pr(attack)	Threshold
1	New York city	New York	8,214,426	0.1101	1.07
2	Los Angeles	California	3,849,378	0.0516	1.45
3	Chicago	Illinois	2,833,321	0.0380	1.60
4	Houston	Texas	2,144,491	0.0287	1.74
5	Phoenix	Arizona	1,512,986	0.0203	1.91
6	Philadelphia	Pennsylvania	1,448,394	0.0194	1.93
7	San Antonio	Texas	1,296,682	0.0174	1.99
8	San Diego	California	1,256,951	0.0168	2.01
9	Dallas	Texas	1,232,940	0.0165	2.01
10	San Jose	California	929,936	0.0125	2.16

- Result: Pr(signal | attack) = 0.388
- Naïve result: Pr(signal | attack) = 0.283

P_d – False Alarm Trade-Off



Choosing γ and κ

- Optimal probability of detection for various choices of γ and κ

$\mathbf{P}_d$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
$\gamma = 1$	0.165	0.228	0.272	0.307	0.336
$\gamma = 2$	0.388	0.481	0.540	0.583	0.618
$\gamma = 3$	0.726	0.801	0.840	0.866	0.885
$\gamma = 4$	0.939	0.964	0.974	0.980	0.984

  - Choice of κ depends on available resources
  - Setting γ is subjective: what size mean increase important to detect?

Sensitivity Analyses

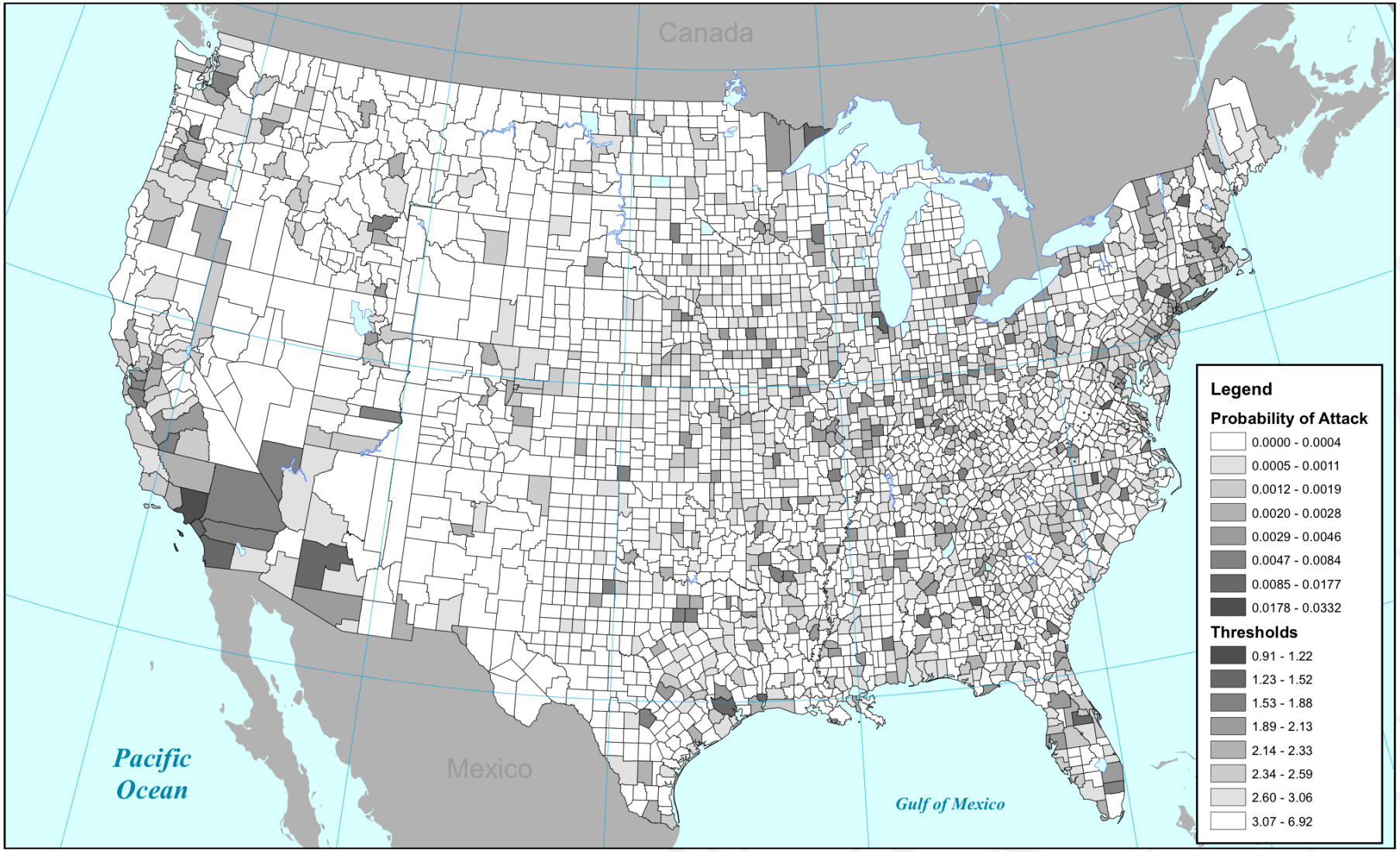
- Optimal probability of detection

$\mathbf{P}_d$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
$\gamma = 1$	0.165	0.228	0.272	0.307	0.336
$\gamma = 2$	0.388	0.481	0.540	0.583	0.618
$\gamma = 3$	0.726	0.801	0.840	0.866	0.885
$\gamma = 4$	0.939	0.964	0.974	0.980	0.984

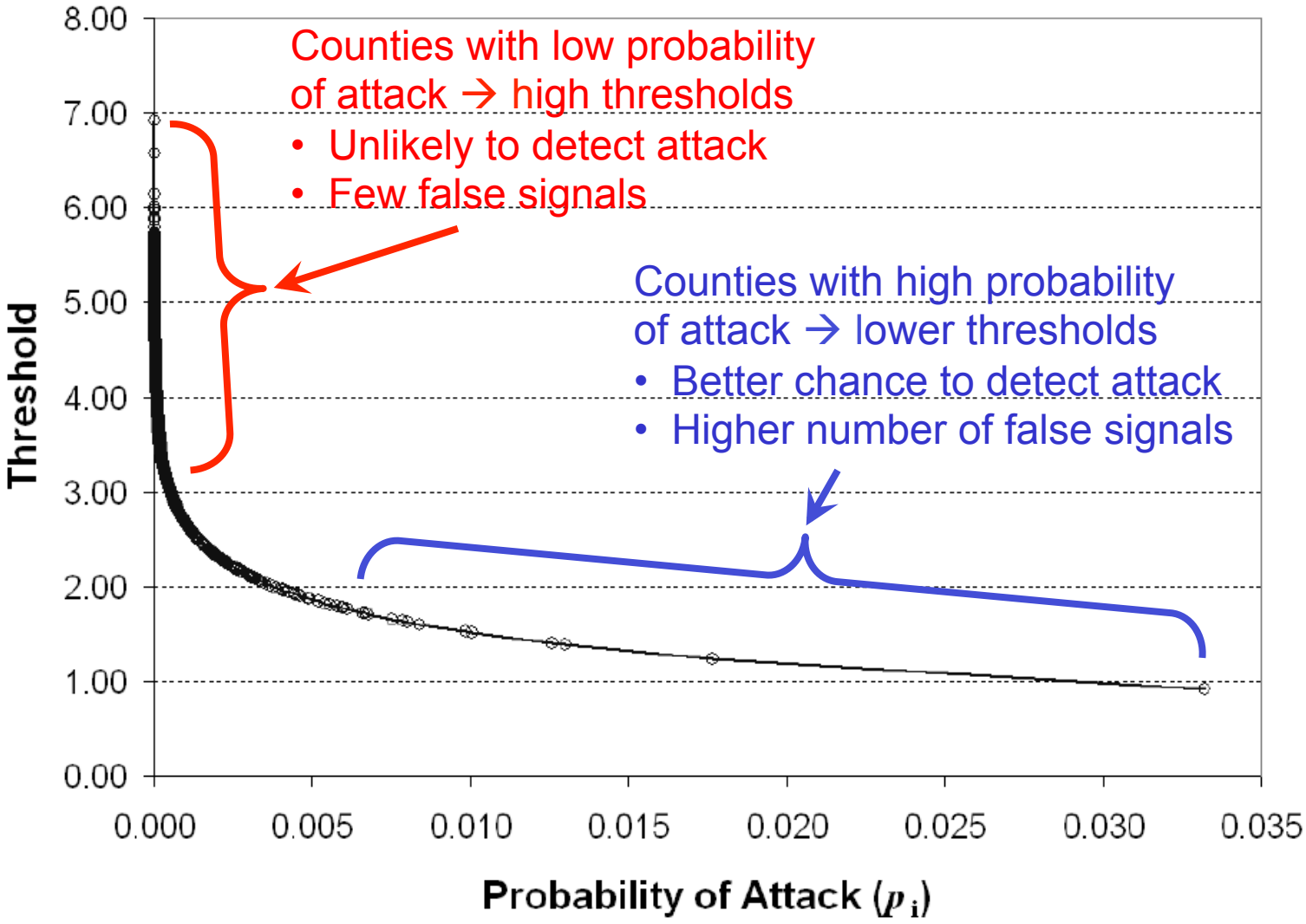
- Actual probability of detection

$\mathbf{P}_d$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
Observed $\gamma = 1$	0.137	0.193	0.235	0.269	0.298
Observed $\gamma = 2$	0.388	0.481	0.540	0.583	0.618
Observed $\gamma = 3$	0.711	0.790	0.832	0.859	0.879
Observed $\gamma = 4$	0.925	0.955	0.968	0.976	0.981

Optimizing a County-level System

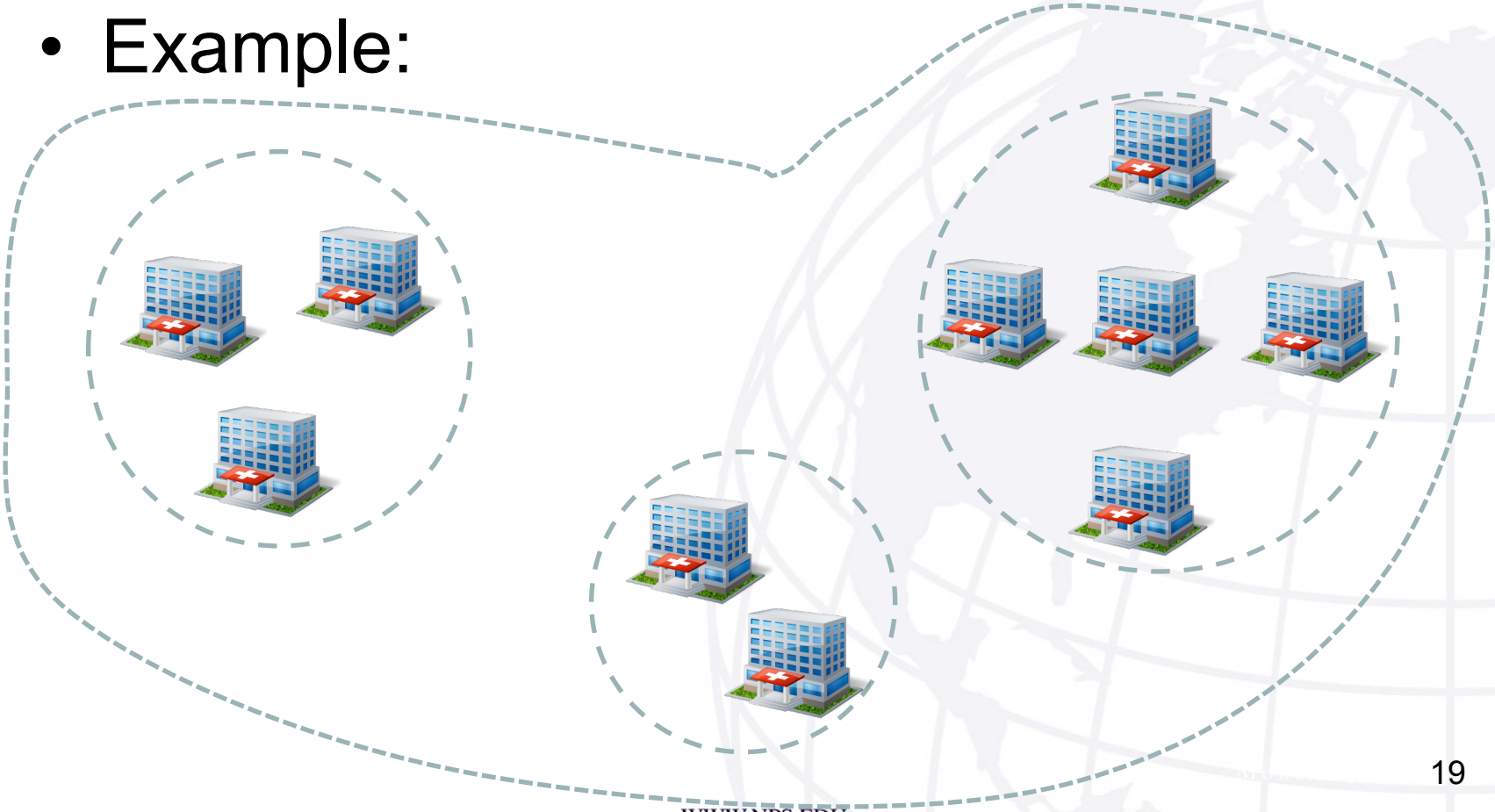


Thresholds as a Function of Probability of Attack



Relaxing the Assumptions

- Some locations may be correlated
  - E.g., hospitals in close proximity
- Example:



Relaxing the Assumptions (cont'd)

- Here $F_{0,i} = N\left(\boldsymbol{\mu}_{0,i}, \boldsymbol{\Sigma}_i\right)$ and $F_{1,i} = N\left(\boldsymbol{\mu}_{1,i}, \boldsymbol{\Sigma}_i\right)$ for i=1,..,k groups, and we'll assume

$$\left\|\boldsymbol{\mu}_{0,i} - \boldsymbol{\mu}_{1,i}\right\| = \nu$$

- For $X_i \sim F_0$

$$\left(\mathbf{X}_{i,t} - \boldsymbol{\mu}_{0,i}\right)\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{X}_{i,t} - \boldsymbol{\mu}_{0,i}\right) \sim \chi^2_{n_i}$$

and for $X_i \sim F_1$

$$\left(\mathbf{X}_{i,t} - \boldsymbol{\mu}_{0,i}\right)\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{X}_{i,t} - \boldsymbol{\mu}_{0,i}\right) \sim \chi^2_{n_i,\nu}$$

- Then, the optimal thresholds are found via

$$\max_{\mathbf{h}} \sum_i \left[1 - \chi^2_{n_i,\nu}(h_i)\right] p_i$$

$$\text{s.t.} \sum_i \left[1 - \chi^2_{n_i}(h_i)\right] \le \kappa$$

Thresholds for An Illustrative Example

Sensor i	Cluster j	p_i	Optimal (Group) Thresholds	"Optimal" (Individual) Thresholds	Adjusted "Optimal" (Individual) Thresholds
1	1	0.45	5.81538	1.8946	1.6094
2		0.05		2.9931	2.7092
3	2	0.20	8.99983	2.2993	2.0150
4		0.05		2.9931	2.7092
5		0.05		2.9931	2.7092
6	3	0.10	13.9231	2.6474	2.3627
7		0.05		2.9931	2.7092
8		0.02		3.4675	3.1674
9		0.02		3.4675	3.1674
10		0.01		3.7725	3.5569
	Specified κ:		0.1	0.1	0.2
	Achieved κ:		0.1	0.06	0.1

Example: Probability of Detection

	Optimal (Group) Thresholds	"Optimal" (Individual) Thresholds	Adjusted "Optimal" (Individual) Thresholds
$\mu_{ij} = 2$ for exactly one i and j	0.46	0.39	0.48
$\mu_{ij} = \sqrt{2}$ for two sensors in cluster j	0.46	0.20	0.28
$\mu_{ij} = \sqrt{4/n_j}$ for all sensors in cluster j	0.46	0.18	0.25

Conclusions

- Can “tune” surveillance networks using intel to improve detection performance
  - Particularly useful for surveillance networks with fixed (immovable) sensors
- Formulation explicitly accounts for allowable false signal rate
  - Failure to do so a major issue with biosurveillance
- More research required to further generalize methods

Other Applications

- Computer intrusion detection
- Terrorist activity detection
- Port or other perimeter security applications

✓ Most generally, monitoring set of data streams with prior information about where anomalies are likely to occur

Selected References

Biosurveillance System Optimization

- Fricker, R.D., Jr., and D. Banschbach (2012). Optimizing Biosurveillance Systems that Use Threshold-based Event Detection Methods, *Information Fusion*, **13**, 117-128.

SPC System Optimization

- Fricker, R.D., Jr. (2009). Optimizing Shewhart Charts in Parallel Production Lines, *International Journal of Quality Engineering and Technology*, **1**, 125-135.

Biosurveillance Background Information

- Fricker, R.D., Jr., *Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance*, to be published by Cambridge University Press. Draft available on-line at http://faculty.nps.edu/rdfricke/OA4910.htm#book.
- Fricker, R.D., Jr. (2011). Some Methodological Issues in Biosurveillance (with commentaries and rejoinder), *Statistics in Medicine*, **30**, 403-441.
- Fricker, R.D., Jr. (2011). Biosurveillance: Detecting, Tracking, and Mitigating the Effects of Natural Disease and Bioterrorism, *Encyclopedia of Operations Research and the Management Sciences*, Cochran, J.J. (ed.), John Wiley & Sons.
- Fricker, R.D., Jr., and H. Rolka (2006). Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4-13.

Selected References (cont'd)

Detection Algorithm Development and Assessment

- Hagen, K.S., R.D. Fricker, Jr., K. Hanni, S. Barnes, and K. Michie (2011). Assessing the Early Aberration Reporting System's Ability to Locally Detect the 2009 Influenza Pandemic, *Statistics, Politics, and Policy*, **2**, issue 1, article 1.
- Fricker, R.D., Jr., Hegler, B.L., and D.A Dunfee (2008). Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms, *Statistics in Medicine,* **27**, pp. 3407-3429.
- Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr. (2008). A One-Sided MEWMA Chart for Health Surveillance, *Quality and Reliability Engineering International,* **24**, pp. 503-519.
- Fricker, R.D., Jr., and J.T. Chang (2008). A Spatio-temporal Method for Real-time Biosurveillance, *Quality Engineering*, **20**, pp. 465-477.
- Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu (2008). Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, *Quality Engineering*, **20**, pp. 478-494.